View Consistent Purification for Accurate Cross-View Localization Supplementary Material

Shan Wang[1,2] Yanhao Zhang[1] Akhil Perincherry[3] Ankit Vora[3] Hongdong Li[1]
[1]Australian National University [2]Data61, CSIRO [3]Ford Motor Company

1. Evaluation of Other FordAV-CVL Dataset Logs

The 'Log4' trajectory was chosen for method evaluation in SIBCL [1] owing to its alignment accuracy with the satellite image. Furthermore, we also evaluated other logs from the FordAV-CVL Dataset in Tab. 1 to supplement the results presented in Tab. 2 of the main paper. There are three travelings included in every log: '2017-08-04-26', '2017-07-24', and '2017-10-26'. For the purpose of training, evaluation, and test dataset split, we use '2017-07-24' as the evaluation dataset for all logs. We select the traveling sequence with a higher number of images as the training dataset. Specifically, the training dataset of 'Log1' and 'Log3' is '2017-10-26', whereas the training dataset of 'Log4' and 'Log5' is '2017-08-04-26'. The results demonstrate that our method is capable of estimating accurate 3-DoF pose with low spatial and angular errors in various scenarios, including freeway (log1), residential (log3, log5), university (log4) and vegetation (log4, log5).

Table 1: Performance of our method on additional logs of the FordAV-CVL dataset

	Lateral						**Longitudinal**						**Yaw**				
	mean↓	median↓	0.25m↑	0.5m↑	1m↑	2m↑	mean↓	median↓	0.25m↑	0.5m↑	1m↑	2m↑	mean↓	median↓	1°↑	2°↑	4°↑
Log1	0.64	0.37	34.98	62.45	84.39	91.94	0.25	0.12	82.86	90.40	95.07	98.47	2.37	0.70	58.84	70.84	81.23
Log3	1.07	0.99	10.64	22.83	50.60	88.89	0.96	0.70	18.79	36.44	65.89	91.22	1.82	1.07	47.86	68.90	87.59
Log5	0.88	0.66	18.36	38.70	70.01	90.30	1.80	0.75	19.82	36.83	58.59	76.76	1.23	0.62	65.51	84.25	93.57
ALL	1.01	0.63	21.35	41.52	66.02	85.01	0.67	0.50	26.56	50.33	80.32	95.00	1.61	0.73	67.20	82.21	88.46

ALL: All images from 'Log1', 'Log3', 'Log4', and 'Log5'. 'Log2' and 'Log6' are excluded due to the construction of road and building during data collection.

2. Performance with Different Initial Poses

In our main paper, we presented a chart illustration in Fig.7. Here, we further provide complete metrics results in Tab. 2 and Tab. 3. By presenting these tables, we aim to provide a comprehensive view of the data and enable readers to analyze the metrics more thoroughly.

Table 2: Performance of our method in different initial pose of the KITTI-CVL dataset

translation	yaw	**Lateral**						**Longitudinal**						**Yaw**				
m	°	mean↓	median↓	0.25m↑	0.5m↑	1m↑	2m↑	mean↓	median↓	0.25m↑	0.5m↑	1m↑	2m↑	mean↓	median↓	1°↑	2°↑	4°↑
15	7.5	0.17	0.12	84.28	98.44	99.42	99.55	0.19	0.17	86.36	99.66	99.66	99.66	4.45	2.03	27.52	49.35	72.70
	10	0.59	0.19	54.94	85.10	96.53	97.20	0.48	0.10	86.44	95.46	96.57	96.71	5.12	2.49	23.61	42.91	64.93
	12.5	2.32	0.20	58.54	79.51	83.57	84.72	2.10	0.18	70.68	81.65	81.71	81.78	6.24	2.70	23.07	41.34	60.64
	15	3.98	0.25	50.25	64.27	67.24	69.28	3.80	0.20	58.45	64.63	64.81	65.06	7.87	3.59	19.89	35.32	52.96
30		0.18	0.13	76.99	96.40	99.89	99.95	0.12	0.10	94.76	99.64	99.88	99.91	5.18	1.92	30.86	51.04	69.59
45	5	0.55	0.25	49.12	78.88	95.27	97.21	0.49	0.21	56.86	88.82	96.22	97.06	8.94	3.16	19.94	36.22	56.45
60		0.45	0.33	39.73	70.11	95.94	99.26	0.39	0.29	44.14	79.55	97.18	99.12	16.04	5.48	15.47	28.21	43.88

In order to facilitate comparison, we have included a performance analysis with a single front onboard camera from the FordAV-CVL dataset, as depicted in Fig. 1.

Table 3: Performance of our method in different initial pose of the FordAV-CVL dataset

translation m	yaw °	Lateral						Longitudinal						Yaw				
		mean↓	median↓	0.25m↑	0.5m↑	1m↑	2m↑	mean↓	median↓	0.25m↑	0.5m↑	1m↑	2m↑	mean↓	median↓	1°↑	2°↑	4°↑
15	7.5	0.66	0.49	26.84	51.07	81.13	97.90	1.25	0.61	23.20	43.30	68.74	85.53	1.22	0.62	68.62	87.22	94.79
	10	0.74	0.50	26.40	50.01	79.61	96.58	1.58	0.64	22.22	41.55	65.80	81.96	1.61	0.64	66.92	85.03	92.79
	12.5	1.09	0.53	25.57	47.92	76.15	93.02	2.15	0.72	20.32	38.35	61.24	76.29	2.84	0.69	63.22	80.65	88.87
	15	1.71	0.58	23.52	44.12	69.72	86.49	3.24	1.02	16.23	30.74	49.69	63.51	4.65	0.86	54.61	71.31	80.57
30		0.60	0.45	27.10	54.47	84.30	97.62	1.03	0.50	25.65	49.85	73.28	88.00	1.04	0.61	71.87	92.41	97.36
45	5	0.72	0.50	26.72	49.96	79.48	96.18	1.21	0.61	22.86	42.70	68.03	85.31	2.11	0.65	66.39	84.53	91.99
60		0.93	0.53	25.39	47.17	74.75	92.12	1.43	0.67	22.01	40.50	64.21	81.03	4.88	0.72	61.51	78.44	85.70

To bring a comprehensive view of the data and enable readers to analyze the metrics more thoroughly, we further provide complete metrics results in Tab. 2 and Tab. 3, as a supplementary of the chat illustration in Fig. 7 in our main paper. In order to facilitate comparison, we have included a performance analysis with a single front onboard camera from the FordAV-CVL dataset, as depicted in Fig. 1.

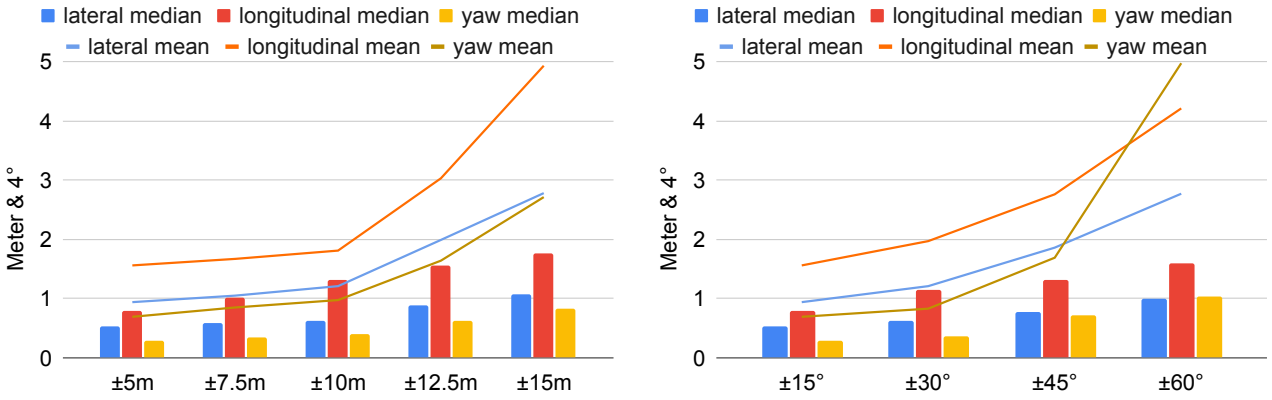

Figure 1: Impact of Initial Pose with one front camera setting in FordAV-CVL dataset. (left) Method performance as initial pose translation varies, with orientation noise fixed within $\pm 15^{\circ}$ range. (right) Method performance as initial pose orientation varies, with translation noise fixed within $\pm 5\mathrm{m}$ range. The vertical axis shows translation error in units of m and orientation error in units of 4°.

3. Visualization of Confidence Maps

The main paper provides an example of the view-consistent confidence map V, the on-ground confidence map O, and the fused confidence map C of the KITTI-CVL dataset. Additionally, we present an example of the FordAV-CVL dataset in Fig. 2. In addition, we have generated confidence map videos of continuous trajectories in both the KITTI-CVL and FordAV-CVL datasets, which can be found in the supplementary video.

References

[1] Shan Wang, Yanhao Zhang, Ankit Vora, Akhil Perincherry, and Hengdong Li. Satellite image based cross-view localization for autonomous vehicle. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3592–3599. IEEE, 2023.

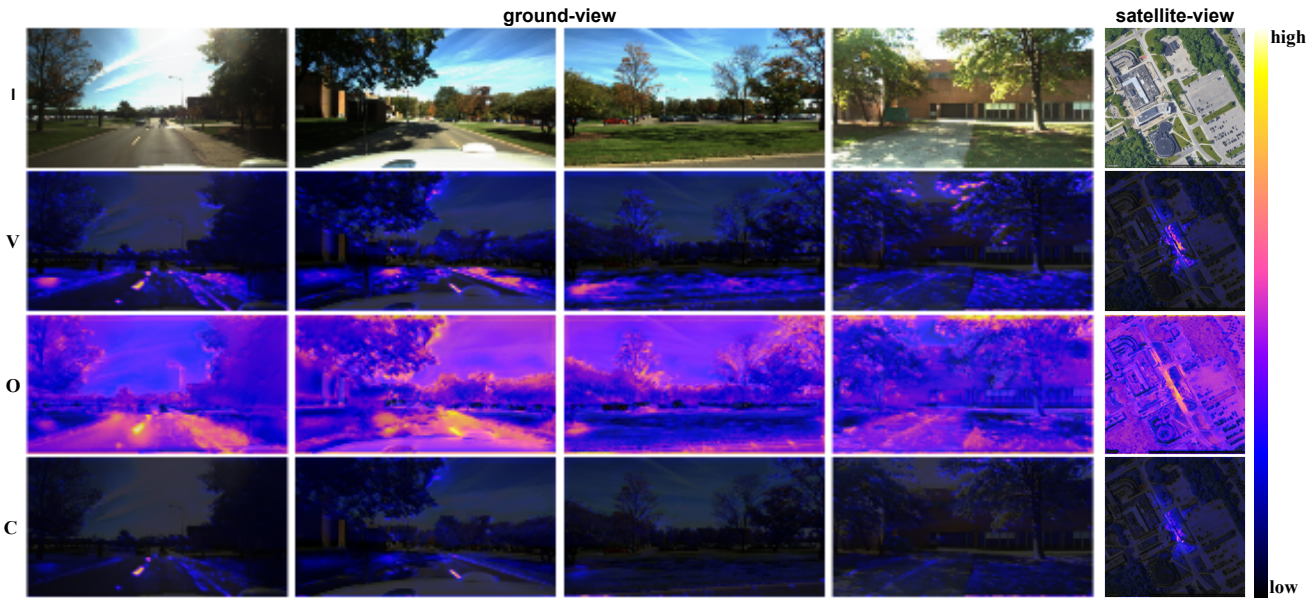

Figure 2: Illustration of confidence maps in FordAV-CVL dataset. The view-consistent confidence map (2^{nd} row) V assigns high confidence to objects that appear consistently in both ground-view and satellite images, such as road marks, and curbs. Conversely, the confidence map assigns low confidence to temporally inconsistent objects, such as vehicles and pedestrians. The on-ground confidence map (3^{rd} row) O highlights only on-ground cues. It is noteworthy that the on-ground confidence map O assigns a high score to the sky and tree leaves. This is because the algorithm only uses key points located under the focal point of ground-view images. Consequently, objects that only exist in the upper part of ground-view images are not supervised and do not affect the localization. The fused confidence map (4^{th} row) C highlights objects that are both view-consistent and on-ground.